Introduction to Statistical Methods in Political Science

Lecture 10: Sampling Distributions for Estimators of Continuous Variables

Ignacio Urbina

Ph.D. Candidate in Political Science

Sampling Distribution of the Sample Mean $\bar{x}$

Our goal is often to estimate the unknown population mean μ using the sample mean $\bar{x}$ calculated from a random sample $X_1, ..., X_n$.

The sample mean $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} X_i$ is itself a random variable, as its value depends on the particular sample drawn.

The **sampling distribution of** $\bar{x}$ describes the probability distribution of the possible values of $\bar{x}$ if we were to repeatedly draw samples of size n from the same population.

Key properties of this distribution are its mean $E(\bar{x})$ and its variance $\text{Var}(\bar{x})$ (or standard error $SE(\bar{x})$).

Case 1: Normal Population (Known σ^2)

Assume the underlying population follows a normal distribution, $X_i \sim N(\mu, \sigma^2)$, and the population variance σ^2 is known.

- The sample mean $\bar{x}$ is **exactly** normally distributed.
- Mean of $\bar{x}$: $E(\bar{x}) = \mu$ (unbiased estimator).
- Variance of $\bar{x}$: $\mathrm{Var}(\bar{x}) = \frac{\sigma^2}{n}$.
- Standard Error of $\bar{x}$: $SE(\bar{x}) = \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}$.

Distribution:

$$\bar{x} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

Standardized Statistic (Z-score):

$$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$$

Case 2: Large Sample Size (CLT)

What if the population distribution is not normal, or unknown?
Central Limit Theorem (CLT): If the sample size n is sufficiently large ($n \geq 30$), the sampling distribution of $\bar{x}$ will be **approximately** normal, regardless of the shape of the population distribution.

- Mean of $\bar{x}$: $E(\bar{x}) = \mu$.
- Variance of $\bar{x}$: $\text{Var}(\bar{x}) = \frac{\sigma^2}{n}$.

Approximate Distribution:

$$\bar{x} \approx N\left(\mu, \frac{\sigma^2}{n}\right) \quad \text{for large } n$$

The CLT is fundamental because it allows us to use normal distribution methods for inference on μ in many practical situations.

The Plug-In Principle (Large Sample)

Usually, the population variance σ^2 is **unknown**.
Plug-In Principle: Estimate σ^2 using the sample variance s^2:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Estimate the standard error of $\bar{x}$ using s:

$$\text{Estimated } SE(\bar{x}) = \frac{s}{\sqrt{n}}$$

For **large samples** ($n \geq 30$), combining CLT and plug-in:

$$Z = \frac{\bar{x} - \mu}{s/\sqrt{n}} \approx N(0,1)$$

This justifies Z-procedures for μ with large samples when σ is unknown.

Case 3: Small Sample Size (Unknown σ^2)

What if n is small ($n < 30$) **and** σ^2 is unknown?
If we assume the population is **normal**, we use the **t-distribution**:

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}} \sim t_{n-1}$$

The t-distribution accounts for the extra uncertainty from estimating σ^2 with s^2.
Details of inference using the t-distribution for small samples will be covered separately.

Why Compare Two Means? Examples

Comparing two sample means helps answer questions across various fields:

- **Business (Job Satisfaction):** Is average job satisfaction (μ_1) in the IT industry different from that in finance (μ_2)? Compare sample means $\bar{x}_1$ and $\bar{x}_2$.
- **Health Science (Physical Activity):** Does a high-intensity exercise regimen (μ_1) lead to a greater mean decrease in cholesterol than a moderate-intensity one (μ_2)? Compare sample mean decreases $\bar{x}_1$ and $\bar{x}_2$.

The goal is to use the sample difference $\bar{x}_1 - \bar{x}_2$ to infer about the population difference $\mu_1 - \mu_2$.

Setup for Comparing Two Means

Consider two **independent** samples:

- Sample 1: Size n_1, mean $\bar{x}_1$, from population with mean μ_1, variance σ_1^2.
- Sample 2: Size n_2, mean $\bar{x}_2$, from population with mean μ_2, variance σ_2^2.

We focus on the sampling distribution of the statistic:

$$\text{Difference in sample means: } \bar{x}_1 - \bar{x}_2$$

Review: Properties of E and Var

Recall fundamental properties: **Expectations (Linearity):**

$$E(aX + bY) = aE(X) + bE(Y)$$

Variances (for Independent X, Y):

$$\text{Var}(aX + bY) = a^2\text{Var}(X) + b^2\text{Var}(Y)$$

These rules are key to deriving the properties of $\bar{x}_1 - \bar{x}_2$.

Expectation of the Difference $\bar{x}_1 - \bar{x}_2$

Using linearity of expectation ($a = 1, b = -1$):

$$E(\bar{x}_1 - \bar{x}_2) = E(\bar{x}_1) - E(\bar{x}_2)$$

Since $E(\bar{x}_1) = \mu_1$ and $E(\bar{x}_2) = \mu_2$:

$$E(\bar{x}_1 - \bar{x}_2) = \mu_1 - \mu_2$$

The difference in sample means is an unbiased estimator of the difference in population means.

Variance and SE of the Difference $\bar{x}_1 - \bar{x}_2$

Assuming the two samples are **independent**: Using the variance rule $(a = 1, b = -1)$:

$$\mathrm{Var}(\bar{x}_1 - \bar{x}_2) = \mathrm{Var}(\bar{x}_1) + \mathrm{Var}(\bar{x}_2)$$

Substitute known variances of sample means:

$$\mathrm{Var}(\bar{x}_1 - \bar{x}_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

The Standard Error (SE) is the square root of the variance:

$$SE(\bar{x}_1 - \bar{x}_2) = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Sampling Distribution of $\bar{x}_1 - \bar{x}_2$ (Large Samples)

If n_1, n_2 are large (CLT), or populations normal (σ's known):

- $\bar{x}_1 \approx N(\mu_1, \sigma_1^2/n_1)$
- $\bar{x}_2 \approx N(\mu_2, \sigma_2^2/n_2)$

Since samples are independent, the difference is also (approx.) normal:

$$\bar{x}_1 - \bar{x}_2 \approx N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$$

This allows Z-procedures for $\mu_1 - \mu_2$ in these cases.

The Plug-In Principle (Two Means, Large Samples)

When σ_1^2, σ_2^2 unknown, but n_1, n_2 large: Estimate SE using sample variances s_1^2, s_2^2:

$$\text{Estimated } SE(\bar{x}_1 - \bar{x}_2) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

The test statistic for $H_0 : \mu_1 - \mu_2 = \Delta_0$ (often $\Delta_0 = 0$):

$$Z = \frac{(\bar{x}_1 - \bar{x}_2) - \Delta_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \approx N(0, 1)$$

Small Samples: t-Distribution (Two Means - Brief Mention)

If either n_1 or n_2 is small, **and** populations assumed normal, **and** σ_1^2, σ_2^2 unknown:
Use the **t-distribution**. The statistic has the form:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Calculating the correct degrees of freedom (df^*) requires specific methods (e.g., Welch-Satterthwaite) unless variances are assumed equal.
Detailed procedures for the two-sample t-test will be covered separately.

Summary and Key Assumptions

- **Sampling Distributions:** Describe the behavior of statistics ($\bar{x}$, $\bar{x}_1 - \bar{x}_2$) over repeated sampling.
- **CLT:** Crucial for large samples, allows using Normal approx. even for non-normal populations.
- **Plug-in Principle:** Use sample variance(s) s^2 when population variance(s) σ^2 are unknown.
- **Independence:** Formulas for $\text{Var}(\bar{x}_1 - \bar{x}_2)$ require independent samples.
- **Large vs. Small Samples:** Use Z-procedures (based on CLT/Normal) for large samples; use t-procedures (based on t-distribution, requires population normality assumption) for small samples when σ's are unknown.
- **Convergence:** For large n, the t-distribution approaches the N(0,1) distribution.